2P-3

RDFを参考にした求人情報メタデータ化方式の提案

中村　大輔†　　福山　峻一†
大阪電気通信大学†

1. はじめに

近年，Linked Open Data（以降，LOD）と呼ばれるWeb間データ共有システムが注目されている．LODは，各分野の公開して意味のある情報を標準的なデータ構造で定義してWeb上に公開し(メタデータ化)，他分野のメタデータと相互リンクすることにより，多様な情報をURIやHTTPの機構を使用して多方面で利用可能にするものである[1]．このためのデータ構造として，RDF(Resource Description Framework)が提唱されている[2]．RDFは，主語，目的語および，これらを関係付ける述語の三つ組みを基本単位としてデータ関の関係を表現する．RDFに準拠してメタデータ化し，LOD化することにより，インターネット上に存在する膨大な情報の統合的再利用の容易化が期待されている．このような期待から公的機関を中心に欧米でLOD化の動きが活発化しており[3]，日本でも国会図書館など件名表目標などでLOD化が始まっている[4]．今後，民間でもLOD化を行っていくためには，より多くの企業や個人が得意分野の情報をメタデータ化してLODとして公開していくことが望まれる．しかし，各種の情報をメタデータ化して公開するための平易な手順が確立されているとは言えない．そこで，本稿では，公開して利用価値が高いと考えられる求人情報をLOD化して公開することを題材に，メタデータ化を行うための汎用的な手順の確立を目的とする．特に，この手順の確立には，データ構造を規定する基本語彙セットの定義方法が中心課題であり，この方法について提案を行う．

2. 基本語彙セット規定のための提案手法

2.1 基本語彙セットの規定プロセス

本稿では，図1に示すプロセスで基本語彙セットの規定を行うものとする．すなわち，まず対象情報領域(求人情報)に存在する用語(語彙候補)を列挙して類似する語をグループ分けする．次に，各グループを代表する語彙を基本語彙セットとして規定する．このようにして規定された基本語彙セットを，RDFにおける述語相当用語として用いることにより，メタデータの記述が可能になる．

2.2 語彙候補の抽出・分類

求人情報項目を分類・抽出するため，大学の就職課に届いている求人票を10社分ランダムに選択して用いる．企業毎にフォーマットは異なるが，その記載項目は企業間での類似性は高い．求人票の情報項目を類似性で分類整理すると，表1の9つのグループ(G1～G9)になる．G1には，求人企業のプロフィール情報を．G2には，募集職種情報を，G3には，休暇，休日，労働時間情報を，G4には，勤務地情報を，G5には企業のPR情報を，G6には応募資格情報を，G7には，賃金情報を，G8には，採用数情報を，そして，G9には，選考情報を，といった具合に各社の求人情報項目を抽出・分類した．これをよく見ると．生起事実を正確に表現するために一般によく用いられている5W1Hに，Whom，How much，How manyを追加した6W3Hにて対応付け可能であると考えられる．このことから，6W3Hの分類軸に着目して，この種情報項目を分類整理し，後述のようにメタデータ化のための基本語彙セットのための語彙候補を抽出する方法論の提案可能性があると考える．

A method for extracting RDF style metadata from job information.
† Daisuke Nakamura, Shunichi Fukuyama, Osaka Electro-Communication University

表1　求人票記載項目の分類例

Group	列挙した求人情報語彙 (語彙候補)	6W3H 要素
G1	求人先，本社所在地，所在地，会社名，代表者，公開日，資本金，年商，売上高，従業員数，HP，URL，e-mail，TEL，株式，業種，事業内容	Who
G2	職務内容，職種，職種名，役職，募集職種	What
G3	フレックスタイム制，コアタイム，標準労働時間，就業時間，勤務時間，休日，週休，休暇，休職，有給休暇，有給休暇数，夏季休暇，年末年始休暇，特別休暇，育児休職/育児休業，介護休職/介護休業，各種特別休暇，リフレッシュ休暇，永年勤続休暇	When
G4	支店等，支店，勤務先予定地，勤務地，勤務地住所，事業所，営業所	Where
G5	会社の特徴，企業紹介，特徴PR社風等，募集理由	Why
G6	応募条件，応募要件，応募資格，短大・専門卒，大卒，募集対象，募集要項，資格等	Whom
G7	給与等，給与，基本給，初任給，賞与，昇給，給与改定，福利厚生，社会保険，加入保険等，労働組合，各種保険等，健康保険，厚生年金，雇用保険，労災保険，財形貯蓄，待遇，手当，諸手当，通勤手当，時間外手当，退職金，慶弔金，従業員持株制度，旅行など	How much
G8	採用予定数，採用人数，募集人数，求人数	How many
G9	選考方法，選考フロー，試験内容，書類提出方法，提出書類，応募受付方法，応募方法，応募締切日，エントリ問い合わせ及び書類提出先	How

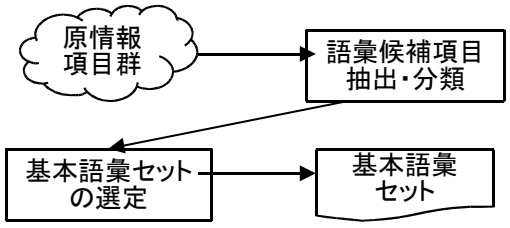

図1　基本語彙セットの抽出プロセス

2.3 求人情報メタデータ化用の語彙選定

メタデータを記述する代表的な語彙セットとしてDublin Core[5]，FOAF[6]などが提案されている．前者は，タイトル，作者，日付など書誌的情報を記述するための語彙セットである．また，幅広く応用可能にするため，それぞれの語彙の定義域や値域は定められていない．後者は，名前，メールアドレスなどの人物の特徴や，知人関係など人に関連する幅広い情報を記述するための語彙セットである．これらの既存語彙を用

いて，求人票情報をメタデータとして記述してみたものが図2である．記述可能な部分はこの程度であり，これらの既存語彙セットは，記述能力面では不十分である．そのため，求人情報をメタデータ化するための新規基本語彙セットを提案する．

表1で列挙した求人票の語彙候補から同義語を代表する語彙を選定して絞込みを行う．同義語を代表用語に一本化する．例としてG1，G3から一部を取り出し，語彙の階層関係を整理すると図3のようになる．

まず，図3の①一世代(親用語が子用語をもたず単一の場合)用語，②二世代(親用語が子用語をもつ場合)にまで展開可能な場合の親用語を基本語彙として選定する．そして，③三世代(親用語が子用語，孫用語を持つ場合)にまで展開可能な場合は，語彙の階層関係上で最上位となる親用語，次の階層で孫用語を持つ子用語を基本語彙として選定する．このとき，基本語彙として選定した親用語と子用語に関連性を持たせ，かつ語彙の意味を精密化する．このため，最上位以外で選定した基本語彙名には，最上位で選定した基本語彙名を必ず含めるようにする．①では資本金をcapital，②では休日をholiday，③では休暇をleave，特別休暇をspecialleave，休職をadministrativeleaveとして基本語彙名を付与した．

次に，選定した基本語彙は既存の法令や慣習に従い，再度適切な表現や定義を与え，最終的に提案すべき基本語彙に入れるべき用語とする．この方法を用いて，全グループについて基本語彙セットを規定したものが表1の下線付きの用語である．また，用語の階層関係や，定義，名前の付け方などの判断は，標準的な業界用語分類などに基づき，改めて識者によるレビューが必要である．

Dublin Coreの場合
dc:creator　(求人先企業名の記述に応用可)
dc:date　(求人情報の公開日の記述に応用可)

FOAFの場合
foaf:homepage　(求人先企業のHPの記述に応用可)
foaf:mbox　(求人先企業のメールアドレスの記述に応用可)

図2　既存語彙を用いたメタデータ記述

①
〈資本金〉　:capital
②
〈休日〉　:holiday
　└─週休
③
〈休暇〉　:leave
　├─有給休暇
　├─夏季休暇
　├─年末年始休暇
　├─〈特別休暇〉:specialleave
　│　├─慶弔休暇
　│　├─リフレッシュ休暇
　│　└─永年勤続休暇
　└─〈休職〉　:administrativeleave
　　├─育児休職
　　└─介護休職

図3　G1，G3の語彙階層関係図

語彙の名前は，メタデータの要素を識別するためのものであり，論理的な意味を担うわけではないが，実際には人間も語彙利用することを考慮すると，できるだけ表意性のある形で命名することが望しい．また，語彙の中では一貫したスタイルで名前付けすることも重要である．加えてRDFでは，語彙名を英小文字で記述するスタイルが一般的とされており，図3のようなアルファベット表現はこれに従ったものである[7]．

3. メタデータとして公開するための実装例

3.1 求人情報項目とRDFの対応付け

LODとしてメタデータを公開するためにはRDFの三つ組み形式に準拠し，データ間の関係を記述する必要がある．主語と目的語の関係を基本語彙(述語)を用いて記述する必要がある．例として，図3で示した基本語彙(capital，holiday，leave)を用いて，A社の資本金額，週休，有給休暇数との関係をRDFの三つ組み表現すると図4のようになる．

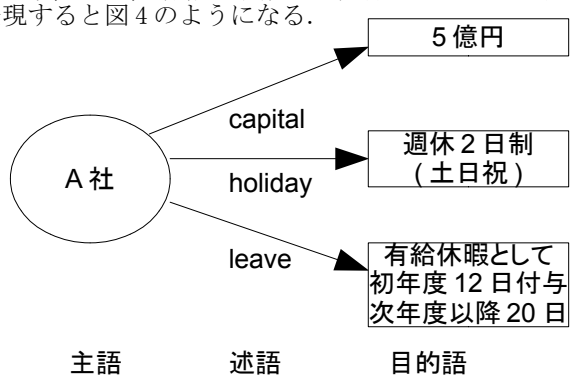

図4　求人情報項目の三つ組み表現例

3.2 RDF/XML形式での実装例

三つ組み表現を最終的にメタデータ化して求人情報を公開する方法としては，RDF形式のデータの作成が必要である．また，RDF/XML[8]形式での記述スタイルが一般的とされている．図4の三つ組み表現をRDF/XMLで記述すると図5のようになる．今回は手書きでRDF/XMLを記述したが，ツールやサービスとして，RDFAuthor[9]，LinkData[10]がある．本格的なメタデータの作成にはこれらを用いることになる．

```
<rdf:Description rdf:about="A社">
  <capital>5億円</capital>
  <holiday>週休2日制(土日祝)</holiday>
  <leave>有給休暇として初年度12日付与 次年度以降20日</leave>
</rdf:Description>
```

図5　図4の三つ組みのRDF/XML表現

4. おわりに

求人情報をメタデータ化する手順として，まず始めに，求人票の情報を抽出・分類した．次に，抽出した情報を元に求人情報をメタデータ化するための基本語彙セットの提案を行った．最後に，提案した基本語彙セットを用いてRDFの三つ組み表現での実装例を示した．今回提案した手法の中で着想した6W3Hは，求人情報のRDFによるメタデータ化のための基本語彙抽出に使用できる．また，6W3H法は他の応用例として，医療カルテ情報，不動産の売買情報のメタデータ化にも応用できる．

参考文献

[1]Bizer, C., et al.: Linked Dataの仕組み Linked Data-The Story So Far, 情報処理, Vol.52, No.3, pp.284-292(2011).
[2]http://www.w3.org/DesignIssues/LinkedData.html
[3]http://richard.cyganiak.de/2007/10/lod/lod-datasets_2010-09-22.html
[4]http://id.ndl.go.jp/auth/ndla
[5]http://www.kanzaki.com/docs/sw/dublin-core.html
[6]http://www.kanzaki.com/docs/sw/foaf.html
[7]神崎 正英 (2005). 『セマンティック・ウェブのための RDF/OWL 入門』 森北出版株式会社
[8]http://www.kanzaki.com/docs/sw/rdf-xml.html
[9]http://rdfweb.org/people/damian/RDFAuthor
[10]http://linkdata.jp/